Regression (IQ) mit KNN

Vorhersage des IQ mit KNN-Regression

In dieser Aufgabe soll mittels KNN-Regression der Intelligenz-Quotient (IQ) von Kindern vorhergesagt werden (siehe auch andere Aufgabe).

KNN-Regression erläutern

k-nearest-neighbors (KNN) ist ein sehr einfaches Verfahren, mit dem Daten klassifiziert oder eine Regression berechnet werden kann. Im Fall der Regression werden für einen Datensatz die nächsten $k \in \mathbb{N}$ Nachbarn ermittelt und der Mittelwert dieser k Response-Variablen berechnet – das ist der geschätzte Wert. ÜblicheANLPrweise wird die euklidische Distanz zum Ermitteln der Nachbarn genutzt.

Ein Einführungsvideo finden Sie unter https://youtu.be/sTJApaBjong. Weitere Infos auch unter https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html

Beschreiben Sie mit eigenen Worten (180 bis 220 Wörter), wie dieses Verfahren für die Regression und Klassifikation angewandt werden kann. Was sind die Vor- und Nachteile? Was sind die Besonderheiten im Vergleich zu anderen Verfahren?

Antwort:

Bei Anwendung des Verfahrens für die Klassifikation wird ein Datensatz mit bekannten Kategorien verwendet und geclustert (z. B. verschiedene Arten von Tumorzellen, die unterschiedliche Merkmale besitzen). Nun soll die Kategorie eines unbekannten Wertes geschätzt werden (z. B. einer unbekannten Zelle). Hierfür werden die nächsten k Datenpunkte betrachtet und jene Kategorie für den unbekannten Wert ausgewählt, denen die meisten der k Nachbarn angehören.

Bei Anwendung des Verfahrens für die Regression, werden für einen unbekannten Wert, wie oben bereits erwähnt, die nächsten k Nachbarn ermittelt, der Mittelwert davon berechnet und darauf basierend ein Wert geschätzt.

Ein Vorteil des KNN-Verfahrens ist, dass das Verfahren leicht zu verstehen und zu implementieren ist und sowohl für Klassifikations- als auch Regressionsprobleme verwendet werden kann. Vor allem für nicht-lineare Daten ist es nützlich, da keine Annahmen über die Daten gemacht werden müssen. Nachteile sind, dass bei größeren Datensätzen die Berechnung der Distanzen zwischen den Datenpunkten sehr aufwendig (teuer) werden kann. Dasselbe Problem tritt bei höher-dimensionalen Daten auf, da es mit mehreren Dimensionen immer aufwendiger wird, die Distanzen zu berechnen. Ein weiterer Nachteil ist, dass die Performance des Verfahrens abhängig von der Wahl für k ist und unterschiedliche Werte für k zu Over- oder Underfitting führen können.

Im Vergleich zu anderen Verfahren ist KNN vor allem nützlich, da keine Annahmen über das Modell gemacht werden müssen, wie z. B. bei der linearen Regression in Aufgabe 1.

Daten importieren

Laden Sie abermals den Datensatz unter https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download in einen Data Frame.

```
data = readRDS(url("https://oc.informatik.hs-mannheim.de/s/K2nJQngd8N6o3M5/download", "rb"))
```

KNN-Regression anwenden

Anwendung

Wenden Sie für diesen Datensatz eine KNN-Regression an, um den IQ von Kindern in Abhängigkeit der Merkmale schätzen zu können. Im Package `caret` gibt es die Lernmethode `knn`, die dafür genutzt werden kann.

```
# Steuerparameter für Modelltraining festlegen
# args: number = wie viele Faltungen sollen in cross validation durchgeführt werden?
# -> Daten werden in 10 Teile aufgeteilt und Modell wird 10 Mal trainiert jeweils mit anderem Testdaten
# repeats: wiederhole cross validation mehrmals
train_control = trainControl(method = "repeatedcv", number = 10, repeats = 3)

# trainiere Modell
knn_model = train(IQ.child ~ IQ.mother + IQ.father + n.siblings,
                  data = data,
                  method = "knn",
                  trControl = train_control,
                  preProcess = c("center", "scale"), # Daten werden zentriert und skaliert
                  tuneLength = 10,
                  kMax = 20)

print(knn_model)
```

```
## k-Nearest Neighbors
##
## 130 samples
##   3 predictor
##
## Pre-processing: centered (3), scaled (3)
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 117, 118, 117, 118, 116, 117, ...
## Resampling results across tuning parameters:
##
##   k   RMSE      Rsquared   MAE
##    5  9.862661  0.5628203  8.008543
##    7  9.549762  0.5827757  7.710460
##    9  9.657625  0.5755676  7.749737
##   11  9.632111  0.5889800  7.714070
##   13  9.693196  0.5875261  7.705505
##   15  9.784860  0.5824648  7.763462
##   17  9.836902  0.5893218  7.842419
##   19  9.886043  0.5950749  7.910188
##   21  9.926645  0.6005511  7.926922
##   23  9.994128  0.6048918  7.940614
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was k = 7.
```

RMSE: Root Mean Squared Error, misst durchschnittliche quadratische Abweichung zwischen vorhergesagten und tatsächlichen Werten Rsquared: wie gut passen vorhergesagte zu tatsächlichen Werten? zwischen 0 und 1, höher -> bessere Modellanpassung MAE: Mean Absolute Error, Durchschnitt der absoluten Werte der Fehler zwischen vorhergesagten und tatsächlichen Werten, niedriger -> höhere Vorhersagengenauigkeit

k?

Finden Sie heraus, welcher Wert für die Anzahl der nächsten Nachbarn am besten funktioniert. Beschreiben Sie, wie Sie das gemacht haben.

k = 7 hat als Wert für die Anzahl der nächsten Nachbarn am besten funktioniert. Laut Zusammenfassung wurde der Wert für k verwendet, bei dem RMSE am niedrigsten ist. RMSE misst die durchschnittliche quadratische Abweichung zwischen den vorhergesagten und tatsächlichen Werten und wird somit als Maß verwendet, wie gut das Verfahren mit dem jeweiligen Wert für k funktioniert.

R^2

Lassen Sie die R^2-Statistik berechnen. Wie sieht diese im Vergleich zur linearen Regression aus? Welches Verfahren ist besser?

```
r_squared_knn = knn_model$results[knn_model$result$k==7, "Rsquared"]

cat("R-squared (KNN):", round(r_squared_knn, 4))
```

```
## R-squared (KNN): 0.5828
```

Der Wert von R^2 für k=7 liegt bei 0.5678. Bei der linearen Regression lag der Wert bei 0.67. Verglichen mit der linearen Regression scheint das KNN-Modell also weniger Variation im IQ der Kinder zu erklären. Um eine Aussage darüber zu treffen, welches Verfahren besser ist, sollten allerdings auch andere Statistiken untersucht werden, wie z. B. Residual-Plots, um die Überprüfung der Annahme der linearen Regression zu beurteilen. Wenn in diesen Plots z. B. Muster oder Strukturen erkennbar sind, könnte das bedeuten, dass die Annahme des Modells nicht richtig, obwohl der R^2-Wert hoch ist.